

Nonextensive information theoretical machine

Chaobing Song, Shu-Tao Xia
Graduate School at Shenzhen, Tsinghua University

Abstract—In this paper, we propose a new discriminative model named *nonextensive information theoretical machine (NITM)* based on nonextensive generalization of Shannon information theory. In NITM, weight parameters are treated as random variables. Tsallis divergence is used to regularize the distribution of weight parameters and maximum unnormalized Tsallis entropy distribution is used to evaluate fitting effect. On the one hand, it is showed that some well-known margin-based loss functions such as $\ell_{0/1}$ loss, hinge loss, squared hinge loss and exponential loss can be unified by unnormalized Tsallis entropy. On the other hand, Gaussian prior regularization is generalized to Student-t prior regularization with similar computational complexity. The model can be solved efficiently by gradient-based convex optimization and its performance is illustrated on standard datasets.

I. INTRODUCTION

As the representatives of statistical learning and ensemble learning methods respectively, support vector machine (SVM) Cortes & Vapnik (1995) and adaboost Freund & Schapire (1997) have got a lot of success in practice. They can both be classified in the margin-based classification methodology Rosset et al. (2004). From the view of loss function, in SVM, hinge loss is employed as measure to find the maximum margin plane. While in adaboost, exponential loss is used to select and combine weak learners. In terms of regularization, ℓ_2 -norm and ℓ_1 -norm corresponds to Gaussian prior and Laplace prior Zhu & Xing (2009) respectively and are often used to control the model complexity of SVM. While in the boosting framework, iterative regularization is often used as approximate ℓ_1 regularization Rosset et al. (2004). In terms of data transform, SVM maps data into high dimension by kernel function, while adaboost transforms data as the output of weak learners.

Two interesting questions are whether we can unify the mathematical form of SVM and adaboost in a common framework and whether loss function, regularization method and data transform method can be expressed by a unified mathematical theory. In this paper, we give an attempt under nonextensive information theory (NIT) framework. In complex systems with long-range interaction, long-time memory and multifractals Tsallis (2001), the equilibrium state often shows power-law distribution instead of exponential distribution. Therefore, the well-known Boltzmann distribution (which is exponential distribution) cannot be well used. NIT as a generalization of Shannon information theory aims to model power-law phenomenon by generalizing Boltzmann-Gibbs-Shannon (BGS) entropy to Tsallis entropy of which the maximum entropy distribution is power-law distribution if the entropy index $q \neq 1$.

In machine learning, there has been some applications of Tsallis entropy and its related concepts such as Tsallis mutual

information kernel Martins et al. (2009), t-logistic regression Ding & Vishwanathan (2010), approximate inference based on t-divergence Ding et al. (2011). In Martins et al. (2009), Tsallis mutual information kernel is proposed by extending Jensen-Shannon divergence and Shannon entropy to Jensen-Tsallis q -difference and Tsallis entropy; in Ding & Vishwanathan (2010), convex loss is extended to nonconvex loss by using q -exponential families; in Ding et al. (2011), approximate inference is used to q -exponential family by defining a new divergence.

Concretely, our contributions are:

- By using the concepts and methods from NIT, we propose nonextensive information theoretical machine (NITM) to address binary classification task. Its solution and explicit primal and dual formulations are given.
- By observation, we show that all the well-known $\ell_{0/1}$ loss, hinge loss, squared hinge loss and exponential loss are the maximum unnormalized Tsallis entropy distribution with different entropy indices q ;
- By using Tsallis divergence and q -expectation, we show that Gaussian prior (ℓ_2 norm) regularization can be extended to the more general Student-t prior regularization with similar computational complexity.
- By considering the existing work of nonextensive mutual information kernel Martins et al. (2009), we show that all the three parts of discriminative model, e.g., loss function, regularization and data transform can be expressed consistently under the framework of NIT.
- By experiments, it is showed that NITM can improve the generalization performance on different standard datasets by tuning entropy indices properly.

II. NONEXTENSIVE INFORMATION THEORY

Nonextensive information theory (NIT) has raised a lot of interest in physical community. In this section, we mainly review some necessary concepts from NIT.

For convenience, firstly q -exponent and q -logarithm Tsallis (2001) are defined as

$$\begin{aligned} \exp_q x &= \begin{cases} (1 + (1 - q)x)_+^{\frac{1}{1-q}}, & q \in \mathbb{R} \setminus \{1\}, \\ \exp x, & q = 1 \end{cases}, \\ \ln_q x &= \begin{cases} \frac{x^{1-q} - 1}{1-q}, & q \in \mathbb{R} \setminus \{1\}, \\ \ln x, & q = 1 \end{cases}, \end{aligned}$$

where $[x]_+$ stands for $\max\{x, 0\}$ and $\exp_1 x = \lim_{q \rightarrow 1} \exp_q x = \exp x$, $\ln_1 x = \lim_{q \rightarrow 1} \ln_q x = \ln x$. By its definition, one has

$$\begin{aligned} \exp_q(\ln_q x) &= x, \\ \ln_q(\exp_q x) &= x. \end{aligned}$$

Corresponding to the definition of exponential family, one can define q -exponential family Amari & Ohara as

$$p(\mathbf{x}; \boldsymbol{\theta}) = \exp_q(\boldsymbol{\theta}^T \mathbf{x} - \psi_q(\boldsymbol{\theta})),$$

where $\boldsymbol{\theta}$ is parameters and $\psi_q(\boldsymbol{\theta})$ is log normalized factor.

In addition, denote indicator function

$$I_\infty(A) = \begin{cases} \infty, & \text{event } A \text{ holds} \\ 0, & \text{else} \end{cases}.$$

Denote the real line and the nonnegative half-line by \mathbb{R} and \mathbb{R}_+ respectively. The set of n -dimensional vectors with positive components of sum 1 is denoted by

$$\Delta_n = \left\{ \mathbf{v}, \mathbf{v} = (v_1, v_2, \dots, v_n)^T \in \mathbb{R}_+^n, \sum_{i=1}^n v_i = 1 \right\}.$$

In addition, denote $\mathbf{1}_n$ as a vector in \mathbb{R}_+^n with all elements 1.

For $\mathbf{p} \in \Delta_n$, Tsallis entropy is defined as Tsallis (1988, 2001)

$$\begin{aligned} S_q(\mathbf{p}) &= k \sum_{i=1}^n p_i \ln_q \frac{1}{p_i} \\ &= \begin{cases} -k \frac{\sum_{i=1}^n p_i^q - 1}{q-1}, & q \in \mathbb{R} \setminus \{1\} \\ -k \sum_{i=1}^n p_i \ln p_i, & q = 1 \end{cases}, \end{aligned}$$

where k is an arbitrary positive constant. For convenience, set $k = 1$ in the following context. For $q = 1$, $S_1(\mathbf{p})$ is equivalent to the definition of Shannon entropy. For $q = 0$ and $i \in \{1, 2, \dots, n\}$, define $p_i^q = 0$ if $p_i = 0$ and $p_i^q = 1$ if $p_i \neq 0$, then

$$S_0(\mathbf{p}) = \|\mathbf{p}\|_0 - 1,$$

where $\|\cdot\|_0$, called ℓ_0 pseudo norm, denotes the number of nonzero elements in vector. If $q < 0$, $S_q(p)$ is convex; if $q > 0$, $S_q(p)$ is concave. In all cases, $S_q \geq 0$ (nonnegativity property). For two independent random variables A and B , with probability mass function $\mathbf{p}_A \in \Delta_{n_A}$ and $\mathbf{p}_B \in \Delta_{n_B}$ respectively, consider the new random variable $A \cup B$ defined by the joint distribution $\mathbf{p}_A \cup \mathbf{p}_B \in \Delta_{n_A n_B}$, then Tsallis (1988),

$$S_q(\mathbf{p}_A \cup \mathbf{p}_B) = S_q(\mathbf{p}_A) + S_q(\mathbf{p}_B) + (1 - q)S_q(\mathbf{p}_A)S_q(\mathbf{p}_B),$$

which is called the nonextensive property of Tsallis entropy. One can immediately see that $q < 1$, $q = 1$ and $q > 1$ respectively correspond to superextensivity (superadditivity), extensivity (additivity) and subextensivity (subadditivity). An axiomatic framework for Tsallis entropy (for all $q \in \mathbb{R}$ and an uniqueness theorem can be seen in dos Santos (1997).

As a measure of similarity on \mathbb{R}_+^n , for $\mathbf{p}, \mathbf{t} \in \mathbb{R}_+^n$ and $q \in \mathbb{R}$, generalized Tsallis divergence Martins et al. (2009) is defined as

$$\begin{aligned} D_q(\mathbf{p} \parallel \mathbf{t}) &= \sum_{i=1}^n -p_i \ln_q \left(\frac{t_i}{p_i} \right) - p_i + t_i \\ &= \begin{cases} \frac{\sum_{i=1}^n p_i^q t_i^{1-q} - qp_i + (q-1)t_i}{q-1}, & q \in \mathbb{R} \setminus \{1\} \\ \sum_{i=1}^n p_i \ln \frac{p_i}{t_i} - p_i + t_i, & q = 1 \end{cases} \quad (1) \end{aligned}$$

For $q = 1$, $D_q(\mathbf{p} \parallel \mathbf{t})$ is the definition of the generalized Kullback-Leibler (KL) divergence Csiszár (1975).

For the case $\mathbf{p}, \mathbf{t} \in \Delta_n$, by the definition of $D_q(\mathbf{p} \parallel \mathbf{t})$ in (1), one has

$$\begin{aligned} D_q(\mathbf{p} \parallel \mathbf{t}) &= \sum_{i=1}^n -p_i \ln_q \left(\frac{t_i}{p_i} \right) \\ &= \begin{cases} \frac{\sum_{i=1}^n p_i^q t_i^{1-q} - 1}{q-1}, & q \in \mathbb{R} \setminus \{1\} \\ \sum_{i=1}^n p_i \ln \frac{p_i}{t_i}, & q = 1 \end{cases}, \end{aligned}$$

which is called the Tsallis divergence on discrete probability distribution. For $q = 1$, $D_1(\mathbf{p} \parallel \mathbf{t})$ is the well-known KL divergence.

Similarly, for two *unnormalized* probability density functions (pdf) $p(\mathbf{x})$ and $t(\mathbf{x})$ on $\mathbf{x} \in \mathbb{R}^n$, the generalized Tsallis divergence can be defined as

$$\begin{aligned} D_q(p(\mathbf{x}) \parallel t(\mathbf{x})) &= \int \left(-p(\mathbf{x}) \ln_q \left(\frac{t(\mathbf{x})}{p(\mathbf{x})} \right) - p(\mathbf{x}) + t(\mathbf{x}) \right) d\mathbf{x} \\ &= \begin{cases} \frac{\int (p^q(\mathbf{x}) t^{1-q}(\mathbf{x}) - qp(\mathbf{x}) + (q-1)t(\mathbf{x})) d\mathbf{x}}{q-1}, & q \in \mathbb{R} \setminus \{1\} \\ \int \left(p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{t(\mathbf{x})} - p(\mathbf{x}) + t(\mathbf{x}) \right) d\mathbf{x}, & q = 1 \end{cases}. \end{aligned}$$

For normalized pdfs $p(\mathbf{x})$ and $t(\mathbf{x})$, where $\int p(\mathbf{x}) d\mathbf{x} = 1$, $\int t(\mathbf{x}) d\mathbf{x} = 1$, one has

$$\begin{aligned} D_q(p(\mathbf{x}) \parallel t(\mathbf{x})) &= \int -p(\mathbf{x}) \ln_q \left(\frac{t(\mathbf{x})}{p(\mathbf{x})} \right) d\mathbf{x} \\ &= \begin{cases} \frac{\int p^q(\mathbf{x}) t^{1-q}(\mathbf{x}) d\mathbf{x} - 1}{q-1}, & q \in \mathbb{R} \setminus \{1\} \\ \int p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{t(\mathbf{x})} d\mathbf{x}, & q = 1 \end{cases}, \end{aligned}$$

which is called the Tsallis divergence on continuous probability distribution. Meanwhile, for the normalized pdf $p(\mathbf{x})$, Tsallis entropy can be defined as

$$S_q(p(\mathbf{x})) = -\frac{\int p^q(\mathbf{x}) d\mathbf{x} - 1}{q-1}.$$

For $q > 0$, $D_q(p(\mathbf{x}) \parallel t(\mathbf{x}))$ is a special case of f-divergence (see Cichocki & Amari (2010) and reference therein), which has the following properties.

- Convexity: $D_q(p(\mathbf{x}) \parallel t(\mathbf{x}))$ is convex with respect to (w.r.t.) both $p(\mathbf{x})$ and $t(\mathbf{x})$;
- Strict Positivity: $D_q(p(\mathbf{x}) \parallel t(\mathbf{x})) \geq 0$ and $D_q(p(\mathbf{x}) \parallel t(\mathbf{x})) = 0$ if and only if $p(\mathbf{x}) = t(\mathbf{x})$.

Because of the two useful properties, the value of $D_q(p(\mathbf{x}) \parallel t(\mathbf{x}))$ with $q > 0$ can be used to measure the similarity between $p(\mathbf{x})$ and $t(\mathbf{x})$. In practice, one can make $p(\mathbf{x})$ get close to $t(\mathbf{x})$ as much as possible by minimizing $D_q(p(\mathbf{x}) \parallel t(\mathbf{x}))$ w.r.t. $p(\mathbf{x})$.

The above two properties also hold for $D_q(\mathbf{p} \parallel \mathbf{t})$ in the discrete case.

Particularly, in the discrete case, let $\mathbf{t} = \frac{1}{n} \mathbf{1}_n$, then for

$\mathbf{p} \in \Delta_n$,

$$\begin{aligned} D_q \left(\mathbf{p} \parallel \frac{1}{n} \mathbf{1}_n \right) &= \sum_{i=1}^n -p_i \ln_q \left(\frac{1}{n} \mathbf{1}_n \right) \\ &= \begin{cases} n^{q-1} \frac{\sum_{i=1}^n p_i^{q-1}}{q-1} - \frac{1-n^{q-1}}{q-1}, & q \in \mathbb{R} \setminus \{1\} \\ \sum_{i=1}^n p_i \ln p_i + \ln n, & q = 1 \end{cases} \\ &= \begin{cases} -n^{q-1} S_q(\mathbf{p}) - \frac{1-n^{q-1}}{q-1}, & q \in \mathbb{R} \setminus \{1\} \\ -S_1(\mathbf{p}) + \ln n, & q = 1 \end{cases}, \end{aligned}$$

which shows that for a fixed q , there exists a one-to-one correspondence between $D_q(\mathbf{p} \parallel \frac{1}{n} \mathbf{1}_n)$ and $S_q(\mathbf{p})$. In fact, the entropy of \mathbf{p} can be understood as the degree of similarity from \mathbf{p} to uniform distribution Shore & Johnson. Therefore, maximizing Tsallis entropy $S_q(\mathbf{p})$ is equivalent to minimizing Tsallis divergence $D_q(\mathbf{p} \parallel \frac{1}{n} \mathbf{1}_n)$.

For unnormalized discrete probability distribution \mathbf{p} and $q > 0$, $D_q(\mathbf{p} \parallel \mathbf{1}_n)$ is also an effective measure to the distance from \mathbf{p} to the unnormalized uniform distribution $\mathbf{1}_n$. Neglecting constants, one can define $-D_q(\mathbf{p} \parallel \mathbf{1}_n)$ as the *unnormalized Tsallis entropy* of \mathbf{p} . Therefore, minimizing $D_q(\mathbf{p} \parallel \mathbf{1}_n)$ can be seen as maximizing unnormalized Tsallis entropy of \mathbf{p} . In order to describe the result in Section III consistently, we define $D_\infty(\mathbf{p} \parallel \mathbf{1}_n)$ by its limit given by

$$\begin{aligned} D_\infty(\mathbf{p} \parallel \mathbf{1}_n) &= \lim_{q \rightarrow +\infty} D_q(\mathbf{p} \parallel \mathbf{1}_n) \\ &= \sum_{i=1}^n -p_i + I_\infty(p_i \leq 1) + n. \end{aligned} \quad (2)$$

III. NONEXTENSIVE INFORMATION THEORETICAL MACHINE

Given a set of instance-label pairs (\mathbf{x}_i, y_i) , $i \in \{1, 2, \dots, m\}$, $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$, $\{\phi_i(\cdot)\}_{i=1}^d$ is a group of fixed basis functions. Denote $\Phi = (\phi_1, \phi_2, \dots, \phi_d) = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m)^T$, where $\phi_j = (\phi_j(\mathbf{x}_1), \phi_j(\mathbf{x}_2), \dots, \phi_j(\mathbf{x}_m))^T$ for $j = 1, 2, \dots, d$ and $\mathbf{f}_i = (\phi_1(\mathbf{x}_i), \phi_2(\mathbf{x}_i), \dots, \phi_d(\mathbf{x}_i))^T$ for $i = 1, 2, \dots, m$. Nonextensive information theoretical machine (NITM) solves the following constrained problem:

$$\min_{\mathbf{p}(\mathbf{w}), \mathbf{z}} D_q(p(\mathbf{w}) \parallel p_0(\mathbf{w})) + C \sum_{i=1}^m \exp_{q'}(-z_i) \quad (3)$$

$$\begin{aligned} \text{s.t. } z_i &= \int y_i \mathbf{f}_i^T \mathbf{w} p^q(\mathbf{w}) d\mathbf{w}, \quad i = 1, 2, \dots, m, \\ \int p(\mathbf{w}) d\mathbf{w} &= 1, \end{aligned} \quad (4)$$

where $\mathbf{w} \in \mathbb{R}^d$ is assumed to be a continuous random vector with normalized pdf $p(\mathbf{w})$. Unlike the common ℓ_2 -norm or ℓ_1 -norm regularization, we impose Bayesian prior $p_0(\mathbf{w})$ on \mathbf{w} and use Tsallis divergence

$$D_q(p(\mathbf{w}) \parallel p_0(\mathbf{w})) = \frac{\int p^q(\mathbf{w}) p_0^{1-q}(\mathbf{w}) d\mathbf{w} - 1}{q-1}$$

to measure the distance of distribution from the posterior distribution $p(\mathbf{w})$ to $p_0(\mathbf{w})$. Instead of using the normal

expectation Zhu & Xing (2009), q -expectation $\int \mathbf{w} p^q(\mathbf{w}) d\mathbf{w}$ in (4) is used Curado & Tsallis (1991). Meanwhile,

$$\exp_{q'}(-z_i) = [1 - (1 - q')z_i]_+^{\frac{1}{1-q'}} \quad (5)$$

can be seen as an unnormalized probability mass distribution (pmf) belonging to q' -exponential family. The sum $\sum_{i=1}^m \exp_{q'}(-z_i)$ is used as loss function. The regularization term and loss function are connected by the constraint of q -expectation (4). $C > 0$ is the regularization parameter to tune the relative weight of the two terms. q and q' are called “entropy indices” in NIT.

Due to the Bayesian-style treatment of \mathbf{w} , the final output used to give a discriminant to a new data \mathbf{x} is the posteriori q -expectation, denoted as

$$\langle \mathbf{w} \rangle_{p^q} = \int \mathbf{w}^T p^q(\mathbf{w}) d\mathbf{w},$$

and the discriminative function is

$$y(\mathbf{x}) = \arg \max_{y \in \{-1, 1\}} y \cdot \mathbf{x}^T \langle \mathbf{w} \rangle_{p^q}.$$

It should be noted that $\langle \mathbf{w} \rangle_{p^q}$ is needed to exist in this paper, but it does not mean the normal expectation $\langle \mathbf{w} \rangle_p$ exists at the same time.

Setting q' to $\{0, 1/2\}$ and taking limit at $q' \rightarrow -\infty, 1$, one has the following result.

Theorem 1. *The well-known $\ell_{0/1}$ loss, hinge loss, squared hinge loss and exponential loss can be unified in q' -exponential family. The corresponding relation with q' can be seen in Table I.*

TABLE I
LOSS FUNCTIONS WITH SPECIFIED q'

q'	$\exp_{q'}(-z)$	Notes
$-\infty$	$I(z < 0)$	$\ell_{0/1}$ loss
0	$[1 - z]_+$	hinge loss
$\frac{1}{2}$	$[1 - \frac{1}{2}z]_+^2$	squared hinge loss
1	$\exp(-z)$	exponential loss

Proof: The proof for $q' = 0, \frac{1}{2}, 1$ is neglected.

For $q' \rightarrow -\infty$, if $z = 0$, then $\exp_{q'}(z) = 1$; if $z > 0$, $[1 - (1 - q')z]_+ = 0$, thus $\exp_{q'}(z) = 0$; if $z < 0$,

$$\begin{aligned} &\lim_{q' \rightarrow -\infty} \ln \exp_{q'}(z) \\ &= \lim_{q' \rightarrow -\infty} \frac{\ln(1 + (1 - q')z)}{1 - q'} \\ &= \lim_{q' \rightarrow -\infty} \frac{z}{1 + (1 - q')z} = 0. \end{aligned}$$

Therefore, if $z < 0$, $\lim_{q' \rightarrow -\infty} \exp_{q'}(z) = 1$. ■

From Theorem 1, $\ell_{0/1}$ loss corresponds to q' -exponential family with $q' \rightarrow -\infty$, which is concave. Hinge loss can be seen as the tightest convex relaxation to $\ell_{0/1}$ loss, which is similar to the relationship between ℓ_1 -norm and ℓ_0 -norm. For $q' = \frac{1}{2}$, the coefficient $\frac{1}{2}$ is only a scale factor and the formulation is equivalent to the standard squared hinge loss $[1 - z]_+^2$ after scaling z . For $q' > 1$, as $z \rightarrow (\frac{1}{1-q'})^+$, $\exp_{q'}(-z) \rightarrow +\infty$. Then if one wants the objective function is bounded in any bounded interval, $q' = 1$, which corresponds to

exponential loss, is the largest value we can choose. Therefore, in this paper, q' is selected in $[0, 1]$.

The general model doesn't constrain the selection of q and $p_0(\mathbf{w})$, but it is necessary to select them carefully for model effectiveness and computational efficiency. In this paper, Student-t distribution is considered, for its good properties.

- Its support is \mathbb{R}^d ;
- By varying its degrees of freedom ν , it can model the heavy tailed distribution with different thickness;
- Taking $\nu \rightarrow +\infty$, it is equivalent to Gaussian distribution;

The general model (3) couples a variational optimization subproblem and a numerical optimization subproblem together. For $q \geq 1$, $D_q(p(\mathbf{w})\|p_0(\mathbf{w}))$ in (3) and q -expectation in (4) are convex w.r.t. $p(\mathbf{w})$. In addition, for $0 \leq q' \leq 1$, $\exp_{q'}(-z_i)$ in (5) is also convex. Therefore, for the entropy indices $q \geq 1$ and $0 \leq q' \leq 1$, the general model is a convex problem w.r.t. $p(\mathbf{w})$ and \mathbf{z} . On the one hand, the problem can be solved directly by some variational optimization technique, or convex optimization method if $D_q(p(\mathbf{w})\|p_0(\mathbf{w}))$ and q -expectation can be explicitly expressed in terms of distribution parameters. On the other hand, one can solve it indirectly by solving the Lagrange dual problem. Our first main result is about the solution of $p(\mathbf{w})$ expressed by Lagrange multipliers and the dual optimization formulation of the general model (3).

Theorem 2. For $q \geq 1$ and $0 \leq q' \leq 1$, the posterior distribution $p(\mathbf{w})$ of the general problem (3) can be expressed in terms of the prior distribution $p_0(\mathbf{w})$ and the Lagrange multipliers as

$$p(\mathbf{w}) = \frac{1}{Z_q(\beta)} p_0(\mathbf{w}) \exp_q(p_0^{q-1}(\mathbf{w}) \beta^T \mathbf{H} \mathbf{w}), \quad (6)$$

where $Z_q(\beta)$ is a normalizable factor, β is the Lagrange multipliers and $\mathbf{H} = (y_1 \mathbf{f}_1, y_2 \mathbf{f}_2, \dots, y_m \mathbf{f}_m)^T$.

Meanwhile, one can solve the primal problem in the dual domain by optimizing the following formulation

$$\begin{aligned} \min_{\beta} \quad & \ln_q(Z_q(\beta)) + CD_{1/q'}(\beta/C \|\mathbf{1}_m) \\ \text{s.t.} \quad & \beta \geq \mathbf{0}. \end{aligned} \quad (7)$$

The posterior distribution $p(\mathbf{w})$ in (6) is parametrized by dual variables β . The factor $p_0^{q-1}(\mathbf{w})$ in $\exp_q(\cdot)$ is emerged by the use of q -expectation, which is the key to get a normalizable solution of $p(\mathbf{w})$. In (7), it shows that minimizing the sum of the unnormalized pmf $\sum_{i=1}^m \exp_{q'}(-z_i)$ w.r.t. \mathbf{z} under the constraint (4) is equivalent to maximizing the unnormalized Tsallis entropy $-D_{1/q'}(\beta/C \|\mathbf{1}_m)$ of the scaled dual variables $\frac{\beta}{C}$ under the nonnegative constraint $\beta > \mathbf{0}$. For $q' \rightarrow 0^+$, i.e., $1/q' \rightarrow +\infty$, according to (2),

$$D_\infty(\beta/C \|\mathbf{1}_m) = \sum_{i=1}^m -\frac{\beta_i}{C} + I_\infty\left(\frac{\beta_i}{C} \leq 1\right) + m, \quad (8)$$

which is equivalent to the dual formulation of hinge loss Zhu & Xing (2009).

In Zhu & Xing (2009), the authors emphasize the advantage of combining maximum entropy learning with maximum margin learning. However, from our perspective, maximum margin

learning is the dual formulation of the maximum unnormalized Tsallis entropy learning. Therefore, maximum entropy learning and maximum margin learning can be unified by the concepts of NIT in the NITM model.

Consider the Student-t prior distribution

$$p_0(\mathbf{w}) = \frac{1}{Z_0} \left(1 + \frac{1}{\nu} \|\mathbf{w}\|_2^2\right)^{-\frac{\nu+d}{2}}, \quad (9)$$

where $Z_0 = \frac{\Gamma(\nu/2) \nu^{d/2} \pi^{d/2}}{\Gamma((\nu+d)/2)}$, d is the dimension of \mathbf{w} , $\nu > 0$ is the degrees of freedom. $\Gamma(\cdot)$ denotes Gamma function. For $\nu > 2$, both the mean and covariance of $p_0(\mathbf{w})$ exist and equal $\mathbf{0}$ and $\frac{\nu}{\nu-2} \mathbf{I}$ respectively.

In order to get an analytic solution, we set

$$\frac{1}{q-1} = \frac{\nu+d}{2} > \frac{d}{2},$$

then

$$q < \frac{2+d}{d}. \quad (10)$$

In addition, if ν is expressed by q , then

$$p_0(\mathbf{w}) = \frac{1}{Z_0} \left(1 + \frac{q-1}{2-d(q-1)} \|\mathbf{w}\|_2^2\right)^{\frac{1}{1-q}}, \quad (11)$$

where Z_0 can be written as

$$Z_0 = \frac{\Gamma\left(\frac{2-d(q-1)}{2(q-1)}\right) \left(\frac{2-d(q-1)}{q-1} \pi\right)^{\frac{d}{2}}}{\Gamma\left(\frac{1}{q-1}\right)}. \quad (12)$$

Imposing the above prior distribution $p_0(\mathbf{w})$, the normalization factor $Z_q(\beta)$ can be expressed explicitly. Thus one has the following concrete results.

Theorem 3. Assume $1 \leq q < \frac{2+d}{d}$ and $0 \leq q' \leq 1$, $\frac{1}{q-1} = \frac{\nu+d}{2}$ and $p_0(\mathbf{w})$ is given in (11). Then the posterior distribution $p(\mathbf{w})$ of the general problem (3) can be expressed in terms of the prior distribution $p_0(\mathbf{w})$ and the Lagrange multipliers as

$$p(\mathbf{w}) = \frac{1}{Z_0 c^{d/2}} \left(1 + \frac{1}{\nu c} \|\mathbf{w} - \boldsymbol{\mu}\|_2^2\right)^{-\frac{\nu+d}{2}}, \quad (13)$$

where

$$\begin{aligned} \boldsymbol{\mu} &= \frac{\nu}{\nu+d} Z_0^{-\frac{2}{\nu+d}} \mathbf{H}^T \beta, \\ c &= 1 - \frac{1}{\nu} \|\boldsymbol{\mu}\|_2^2, \end{aligned} \quad (14)$$

where Z_0 is given in (12). For convenience, ν is used in the above formulation.

Meanwhile, one can solve the primal problem in the dual domain by optimizing the following formulation

$$\begin{aligned} \min_{\beta} \quad & \ln_q \left(\exp_q^r \left(\frac{r}{2} Z_0^{2(1-q)} \|\mathbf{H}^T \beta\|_2^2 \right) \right) \\ & + CD_{1/q'}(\beta/C \|\mathbf{1}_m) \\ \text{s.t.} \quad & \beta \geq \mathbf{0}, \end{aligned} \quad (15)$$

where $r = \frac{2+d(1-q)}{2}$, $\mathbf{H} = (y_1 \mathbf{f}_1, y_2 \mathbf{f}_2, \dots, y_m \mathbf{f}_m)^T$ and Z_0 is given in (12). For $q = 1$, it becomes the following ℓ_2 -norm

regularized problem

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{2} \|\mathbf{H}^T \beta\|_2^2 + C D_{1/q'}(\beta/C \|\mathbf{1}_m\|) \\ \text{s.t.} \quad & \beta \geq \mathbf{0}. \end{aligned}$$

Similar to $p_0(\mathbf{w})$, $p(\mathbf{w})$ in (13) is also a Student-t distribution. The variance c is decided uniquely by μ and ν . Optimizing $p(\mathbf{w})$ is equivalent to updating the parameters μ of $p_0(\mathbf{w})$ according to (14), which generalizes the conjugate prior property of exponential family. In the dual formulation (15), one can see that (15) generalizes the dual formulation of ℓ_2 -norm regularizer by imposing an outer function on $\|\mathbf{H}^T \beta\|_2^2$.

Based on the solution (13) of $p(\mathbf{w})$, one can also solve the primal problem directly by simplifying $D_q(p(\mathbf{w})\|p_0(\mathbf{w}))$. For simplicity, we use ν instead of q in the following results.

Theorem 4. For $\nu > 0$ and $0 \leq q' \leq 1$, $p_0(\mathbf{w})$ and $p(\mathbf{w})$ are given in (9), (13) respectively. One can also directly solve NITM by optimizing the following problem

$$\begin{aligned} \min_{\mu} \quad & \frac{1}{2} \left(1 - \frac{1}{\nu} \|\mu\|_2^2 \right)^{-\frac{d}{\nu+d}} \left(\frac{\nu-d}{\nu} \|\mu\|_2^2 + \nu + d \right) \\ & - \frac{\nu+d}{2} + C \sum_{i=1}^m \exp_{q'}(-z_i) \\ \text{s.t.} \quad & z_i = \frac{\nu^{\frac{\nu}{\nu+d}} \pi^{-\frac{d}{\nu+d}} \left(\frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})} \right)^{\frac{2}{\nu+d}}}{\nu+d} \\ & \cdot \left(1 - \frac{1}{\nu} \|\mu\|_2^2 \right)^{-\frac{d}{\nu+d}} y_i \mathbf{f}_i^T \mu, \\ & \text{for } i = 1, 2, \dots, m, \end{aligned} \quad (16)$$

where μ is the posterior expectation of \mathbf{w} .

From Theorem 4, minimizing Tsallis divergence from $p(\mathbf{w})$ to $p_0(\mathbf{w})$ w.r.t. $p(\mathbf{w})$ is equivalent to minimizing a convex numerical optimization problem w.r.t. μ .

Summarizing the above results, NITM unifies $\ell_{0/1}$ loss, hinge loss, squared hinge loss and exponential loss by *unnormalized Tsallis entropy* with single parameter q' . Meanwhile, NITM unifies Gaussian prior and Student-t prior by *Tsallis divergence* and q -expectation with single parameter q . Furthermore, NITM unifies loss function and regularization by the concepts of NIT. In Martins et al. (2009), the authors showed nonextensive information theory can also be used in the design of kernel, named *Tsallis mutual information kernel*. By this framework, they unify the existed linear kernel, Jensen-Shannon kernel and boolean kernel in one parametric family. Therefore, we have Proposition 1.

Proposition 1. All the three parts loss function, regularization and data transform of discriminant model can be described consistently by nonextensive information theory.

Unlike MaxEnDNet in Zhu & Xing (2009) which needs resort to variational approximation, we can directly optimize the dual formulation (15) or the primal formulation (16) based on gradient-based convex optimization. After optimizing β in (15) or μ in (16), the posterior distribution $p(\mathbf{w})$ can be acquired in (13).

IV. EXPERIMENTS

We illustrate the performance of NITM with Student-t prior (11) on standard datasets. The concrete settings are

- 6 standard datasets: appendicitis, australian, banana, hepatitis, ionosphere, magic¹. Each dataset is divided into 10 parts by distribution optimally balanced stratified cross-validation (DOB-SCV) (see Moreno-Torres et al. (2012) and reference therein). 3 parts of them are used as test dataset, while the other 7 parts are used in cross validation.
- Feature transform: for nominal features, we transform them into double values according to their number which starts from 1. Before learning, all the features are normalized with 0 mean and unit length. In addition, a column with all 1 are added to the feature matrix to learn a bias parameter. In this paper, the main interest is the influence of regularization and loss function to empirical generalization performance, so the group of basis functions $\{\phi_i(\cdot)\}_{i=1}^d$ are set as identity matrix.
- Parameter setting: NITM has 3 parameters, ν , q' and C . Since NITM includes the existing hinge loss-based SVM, squared hinge loss-based SVM and exponential loss-based classifier as special cases, in this paper NITM is treated as a meta model. Instances of NITM with concrete values of pair (ν, q') are treated as different models. Meanwhile, C is treated as an inner hyperparameter of model. For an instance of NITM with given (ν, q') , C is selected by 7-cross validation on the divided 7 parts of each dataset. Then instances of NITM with selected C are compared by test error on the rest uninfluenced 3 parts. In experiments, we compare 66 models with ν from $\{1, 10, 10^2, 10^3, 10^4, +\infty\}$ and q' from $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. The inner hyperparameter C of each model is selected among $\{1, 10^2, 10^4, 10^6, 10^8, 10^{10}\}$.
- Algorithms: In experiments, we mainly explore the primal convex optimization method to solve NITM. For the model with $q' > 0$, the optimization problem in (16) is smooth, and thus BFGS method is employed. For $q' = 0$, which corresponds to hinge loss, the optimization problem is nonsmooth, therefore subgradient BFGS method Yu et al. (2008) is employed. In addition, backtracking line search is used to get global solution and speeds up the iteration. For each problem, the iteration will be stopped if the number of iterations exceeds 5000 or the direction vector is orthogonal with gradient vector.
- Result representation: The result is represented in Fig. 1. Each subfigure corresponds to a dataset and reflects the test error as q' changes. It deserves to note that for each pair (ν, q') , C has been selected in the cross validation stage, so the parameter C of each curve is different in general. The legend on the upper left subfigure is shared among the 6 subfigures.

In Fig. 1, it is showed that the curves in each subfigure are quite different, which reflects the different physical character-

¹Available at <http://keel.es/datasets.php>

istics of datasets. In order to explain the role of ν, q' , the result is analyzed by the order of datasets.

- **Appendicitis:** It is showed that better performance is acquired when q' is relatively small ($q \in \{0, 0.1, 0.2, 0.3, 0.4\}$). However, in general, ν has little influence on test error, except for quite small $\nu = 1$, which get smaller test error for many $q' \in \{0.6, 0.7, 0.8\}$ comparing with other values of ν .
- **Australian:** In general, test error will be small if q' is large. For $q' = 0.6$, the best performance is acquired. For fixed q' , a large ν is preferred.
- **Banana:** The best result will be got when $q' = 0.1$. Meanwhile, although ν has little influence, the best test error is got when $\nu = 1$.
- **Hepatitis:** This dataset prefer middle value of ν , e.g., $\nu = 10$. The test error will be 0 if $\nu = 10, q' = 0.3$. In addition, the curve with $\nu = 1$ has different shape from that with other ν 's.
- **Ionosphere:** The consistent shape of the 6 curves shows that this dataset prefer small q' and large ν . The best result is got in the case with $q' = 0, \nu \rightarrow +\infty$, which corresponds to standard hinge loss-based SVM.
- **Magic:** This consistent shape shows that large q' and large ν is preferred. Then best result is acquired when $q' = 1.0, \nu \rightarrow +\infty$, which corresponds to exponential loss with ℓ_2 regularization.

The result shows that different datasets prefer different settings of (ν, q') which is a verification of no-free lunch theorem Wolpert (2002). Although the result seems to be disorder, it is showed that compared with only tuning C , tuning q', ν, C independently is not equivalent to tuning C only and can give extra gain of generalization performance.

V. PROOFS

A. Proof of Theorem 2

Proof: As we say, for $q \geq 1$ and $0 \leq q' \leq 1$, the general problem is a convex program. The Lagrangian associated with the general model is

$$\begin{aligned} & \mathcal{L}(p(\mathbf{w}), \mathbf{z}, \beta_0, \beta) \\ = & D_q(p(\mathbf{w}) \| p_0(\mathbf{w})) + C \sum_{i=1}^m [1 - (1 - q')z_i]_+^{\frac{1}{1-q'}} \\ & + \beta_0 \left(\int p(\mathbf{w}) d\mathbf{w} - 1 \right) \\ & + \sum_{i=1}^m \beta_i \left(z_i - \int y_i \mathbf{f}_i^T \mathbf{w} p^q(\mathbf{w}) d\mathbf{w} \right) \end{aligned}$$

The Lagrangian dual function is defined as $\mathcal{L}^*(\beta_0, \beta) = \inf_{p(\mathbf{w}), \mathbf{z}} \mathcal{L}(p(\mathbf{w}), \mathbf{z}, \beta_0, \beta)$. Denote $\mathbf{H} = (y_1 \mathbf{f}_1, y_2 \mathbf{f}_2, \dots, y_m \mathbf{f}_m)^T$. For $q > 1$, taking the variational derivative of \mathcal{L} w.r.t. p , one gets

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{q}{q-1} \left(\frac{p}{p_0} \right)^{q-1} + \beta_0 - \beta^T \mathbf{H} \mathbf{w} \cdot q p^{q-1}$$

Setting the variational derivative to 0, one has the following expression,

$$\begin{aligned} p(\mathbf{w}) &= \frac{1}{Z_q} \left[p_0^{1-q}(\mathbf{w}) + (1-q) \beta^T \mathbf{H} \mathbf{w} \right]_+^{\frac{1}{1-q}}, \\ &= \frac{1}{Z_q} p_0(\mathbf{w}) \exp_q(p_0^{q-1}(\mathbf{w}) \beta^T \mathbf{H} \mathbf{w})^{\frac{1}{1-q}} \end{aligned}$$

which uses Tsallis cut-off prescription Teweldeberhan et al. (2005) for $1 + (1-q)p_0^{q-1}(\mathbf{w}) \beta^T \mathbf{H} \mathbf{w} < 0$ and $Z_q = \int p_0(\mathbf{w}) \exp_q(p_0^{q-1}(\mathbf{w}) \beta^T \mathbf{H} \mathbf{w})^{\frac{1}{1-q}} d\mathbf{w}$ is a normalization constant and $\beta_0 = \frac{q Z_q^{q-1}}{1-q}$.

For $q = 1$, similarly one gets

$$\frac{\partial \mathcal{L}}{\partial p} = 1 + \ln \frac{p}{p_0} + \beta_0 - \beta^T \mathbf{H} \mathbf{w}$$

Setting the derivative to 0, one has

$$p(\mathbf{w}) = \frac{1}{Z_1} p_0(\mathbf{w}) \exp(\beta^T \mathbf{H} \mathbf{w}),$$

where $Z_1 = \int p_0(\mathbf{w}) \exp(\beta^T \mathbf{H} \mathbf{w}) d\mathbf{w}$ and $\beta_0 = -1 + \ln Z_1$.

For $0 < q' < 1$, substituting $p(\mathbf{w})$ and β_0 into \mathcal{L} , one has

$$\begin{aligned} & \mathcal{L}^*(\beta) \\ = & \inf_{p(\mathbf{w}), \mathbf{z}} \mathcal{L}(p(\mathbf{w}), \mathbf{z}, \beta_0, \beta) \\ = & -\ln_q(Z_q(\beta)) + C \frac{\sum_{i=1}^m (q'(\frac{\beta_i}{C})^{\frac{1}{q'}} - \frac{\beta_i}{C})}{q' - 1} \\ & + \sum_{i=1}^m I_\infty(\beta_i \geq 0), \\ = & -\ln_q(Z_q(\beta)) - C D_{1/q'}(\beta/C \| \mathbf{1}_m) + C m \end{aligned}$$

Similarly, for $q' \rightarrow 0$ and $q' \rightarrow 1$,

$$\begin{aligned} & \mathcal{L}^*(\beta) \\ = & \inf_{p(\mathbf{w}), \mathbf{z}} \mathcal{L}(p(\mathbf{w}), \mathbf{z}, \beta_0, \beta) \\ = & -\ln_q(Z_q(\beta)) + \sum_{i=0}^m \beta_i \\ & + \sum_{i=1}^m I_\infty(0 \leq \beta_i \leq C), \\ = & -\ln_q(Z_q(\beta)) - C D_\infty(\beta/C \| \mathbf{1}_m) + C m \end{aligned}$$

and

$$\begin{aligned} & \mathcal{L}^*(\beta) \\ = & \inf_{p(\mathbf{w}), \mathbf{z}} \mathcal{L}(p(\mathbf{w}), \mathbf{z}, \beta_0, \beta) \\ = & -\ln_q(Z_q(\beta)) + C \sum_{i=1}^m \left(\frac{\beta_i}{C} - \frac{\beta_i}{C} \ln \frac{\beta_i}{C} \right) \\ & + \sum_{i=1}^m I_\infty(\beta_i \geq 0), \\ = & -\ln_q(Z_q(\beta)) - C D_1(\beta/C \| \mathbf{1}_m) + C m, \end{aligned}$$

where $I(\cdot)$ is an indicator function defined in Section 2. Neglecting constant Cm , Theorem 2 is proved. ■

B. Proof of Theorem 3

Proof: Impose the prior distribution (9) and set $\frac{1}{q-1} = \frac{\nu+d}{2}$, then

$$\begin{aligned} & p_0^{1-q}(\mathbf{w}) + (1-q)\beta^T \mathbf{H}\mathbf{w} \\ &= \frac{1}{Z_0^{1-q}} \left(1 + \frac{1}{\nu} \mathbf{w}^T \mathbf{w} \right) - \frac{2}{\nu+d} \beta^T \mathbf{H}\mathbf{w} \\ &= \frac{1}{Z_0^{1-q}} \left(c + \frac{1}{\nu} \|\mathbf{w} - \boldsymbol{\mu}\|_2^2 \right) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu} &= \frac{\nu}{\nu+d} Z_0^{-\frac{2}{\nu+d}} \mathbf{H}^T \boldsymbol{\beta}, \\ c &= 1 - \frac{1}{\nu} \|\boldsymbol{\mu}\|_2^2 \\ &= 1 - \frac{\nu}{(\nu+d)^2} Z_0^{-\frac{4}{\nu+d}} \|\mathbf{H}^T \boldsymbol{\beta}\|_2^2. \end{aligned}$$

Then if $c < 0$,

$$\begin{aligned} & p(\mathbf{w}) \\ &= \frac{1}{Z_q} \left[p_0^{1-q} + (1-q)\beta^T \mathbf{H}\mathbf{w} \right]_+^{\frac{1}{1-q}} \\ &= \frac{1}{Z_q Z_0} (-c)^{\frac{1}{1-q}} \left[\frac{1}{-\nu c} \|\mathbf{w} - \boldsymbol{\mu}\|_2^2 - 1 \right]_+^{\frac{1}{1-q}} \end{aligned}$$

By our setting, $\frac{1}{1-q} = -\frac{\nu+d}{2} \leq -\frac{d}{2} < -\frac{1}{2}$. Then $p(\mathbf{w})$ is unnormalizable and do not satisfy the constraint $\int p(\mathbf{w}) d\mathbf{w} = 1$. Similarly, if $c = 0$, $p(\mathbf{w})$ is also not unnormalizable. Therefore, in our setting, $c > 0$. Then we have

$$p(\mathbf{w}) = \frac{1}{Z_q Z_0} c^{\frac{1}{1-q}} \left[1 + \frac{1}{\nu c} \|\mathbf{w} - \boldsymbol{\mu}\|_2^2 \right]^{\frac{1}{1-q}}$$

From the fact

$$\int \frac{1}{Z_0 c^{d/2}} \left[1 + \frac{1}{\nu c} \|\mathbf{w} - \boldsymbol{\mu}\|_2^2 \right]^{\frac{1}{1-q}} d\mathbf{w} = 1.$$

and $\int p(\mathbf{w}) d\mathbf{w} = 1$, it follows that

$$\begin{aligned} Z_q &= c^{\frac{1}{1-q} + \frac{d}{2}} = c^{-\frac{\nu}{2}} \\ &= \left(1 - \frac{\nu}{(\nu+d)^2} Z_0^{-\frac{4}{\nu+d}} \|\mathbf{H}^T \boldsymbol{\beta}\|_2^2 \right)^{-\frac{\nu}{2}} \\ &= \exp_q^r \left(\frac{r}{2} Z_0^{2(1-q)} \|\mathbf{H}^T \boldsymbol{\beta}\|_2^2 \right), \end{aligned} \quad (17)$$

where $r = \frac{2+d(1-q)}{2}$, Z_0 is given in (12).

Substituting (17) into Theorem 2 and simplifying the case $q \rightarrow 1(\nu \rightarrow +\infty)$, Theorem 3 is proved. ■

C. Proof of Theorem 4

Proof: From the Proof of Theorem 3,

$$p(\mathbf{w}) = \frac{1}{Z_0 c^{d/2}} \left[1 + \frac{1}{(\nu-2)c} \|\mathbf{w} - \boldsymbol{\mu}\|_2^2 \right]^{\frac{1}{1-q}},$$

where

$$c = 1 - \frac{1}{\nu-2} \|\boldsymbol{\mu}\|_2^2. \quad (18)$$

Use the definition of Tsallis divergence and $\frac{1}{q-1} = \frac{\nu+d}{2}$, we can get

$$\begin{aligned} & D_q(p(\mathbf{w}) \| p_0(\mathbf{w})) \\ &= \frac{1}{2} \left(1 - \frac{1}{\nu} \|\boldsymbol{\mu}\|_2^2 \right)^{-\frac{d}{\nu+d}} \left(\frac{\nu-d}{\nu} \|\boldsymbol{\mu}\|_2^2 + \nu + d \right) - \frac{\nu+d}{2} \end{aligned}$$

Use the formulation of normalized Student t distribution, one can compute the constraint (4) as

$$\begin{aligned} z_i &= \frac{\nu^{\frac{\nu}{\nu+d}} \pi^{-\frac{d}{\nu+d}} \left(\frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})} \right)^{\frac{2}{\nu+d}}}{\nu + d} \\ &\quad \cdot \left(1 - \frac{1}{\nu} \|\boldsymbol{\mu}\|_2^2 \right)^{-\frac{d}{\nu+d}} y_i \mathbf{f}_i^T \boldsymbol{\mu}, \\ &\quad \text{for } i = 1, 2, \dots, m. \end{aligned}$$

Substituting it into the general model 3, Theorem 4 is proved. ■

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new discriminant model named nonextensive information theoretical machine (NITM) based on nonextensive information theory. NITM gives a consistent view of regularization and loss function and takes $\ell_{0/1}$, hinge loss, squared hinge loss and exponential loss as special cases. The solution and explicit primal and dual formulations are given. Then experiments show the improvement of generalization performance by tuning ν, q' .

REFERENCES

- Amari, Shun-ichi and Ohara, Atsumi. Geometry of q-exponential family of probability distributions. *Entropy*, 13: 1170–1185.
- Cichocki, Andrzej and Amari, Shun-ichi. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.
- Cortes, Corinna and Vapnik, Vladimir. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- Csiszár, Imre. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pp. 146–158, 1975.
- Curado, Evaldo MF and Tsallis, Constantino. Generalized statistical mechanics: connection with thermodynamics. *Journal of Physics A: Mathematical and General*, 24(2):L69, 1991.
- Ding, Nan and Vishwanathan, SVN. t-logistic regression. In *Advances in Neural Information Processing Systems*, pp. 514–522, 2010.
- Ding, Nan, Qi, Yuan, and Vishwanathan, Svn. t-divergence based approximate inference. In *Advances in Neural Information Processing Systems*, pp. 1494–1502, 2011.
- dos Santos, Roberto JV. Generalization of shannon theorem for tsallis entropy. *Journal of Mathematical Physics*, 38(8): 4104, 1997.
- Freund, Yoav and Schapire, Robert E. A decision-theoretic generalization of on-line learning and an application to

- boosting. *Journal of computer and system sciences*, 55(1): 119–139, 1997.
- Martins, André FT, Smith, Noah A, Xing, Eric P, Aguiar, Pedro MQ, and Figueiredo, Mário AT. Nonextensive information theoretic kernels on measures. *The Journal of Machine Learning Research*, 10:935–975, 2009.
- Moreno-Torres, Jose G, Sáez, José A, and Herrera, Francisco. Study on the impact of partition-induced dataset shift on-fold cross-validation. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(8):1304–1312, 2012.
- Rosset, Saharon, Zhu, Ji, and Hastie, Trevor. Boosting as a regularized path to a maximum margin classifier. *The Journal of Machine Learning Research*, 5:941–973, 2004.
- Shore, John E. and Johnson, Rodney W. axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE transactions on information theory*, 1:26–37.
- Teweldeberhan, AM, Plastino, AR, and Miller, HG. On the cut-off prescriptions associated with power-law generalized thermostatics. *Physics Letters A*, 343(1):71–78, 2005.
- Tsallis, Constantino. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1-2):479–487, 1988.
- Tsallis, Constantino. I. nonextensive statistical mechanics and thermodynamics: Historical background and present status. In *Nonextensive statistical mechanics and its applications*, pp. 3–98. Springer, 2001.
- Wolpert, David H. The supervised learning no-free-lunch theorems. In *Soft Computing and Industry*, pp. 25–42. Springer, 2002.
- Yu, Jin, Vishwanathan, SVN, Günter, Simon, and Schraudolph, Nicol N. A quasi-newton approach to non-smooth convex optimization. In *Proceedings of the 25th international conference on Machine learning*, pp. 1216–1223. ACM, 2008.
- Zhu, Jun and Xing, Eric P. Maximum entropy discrimination markov networks. *The Journal of Machine Learning Research*, 10:2531–2569, 2009.

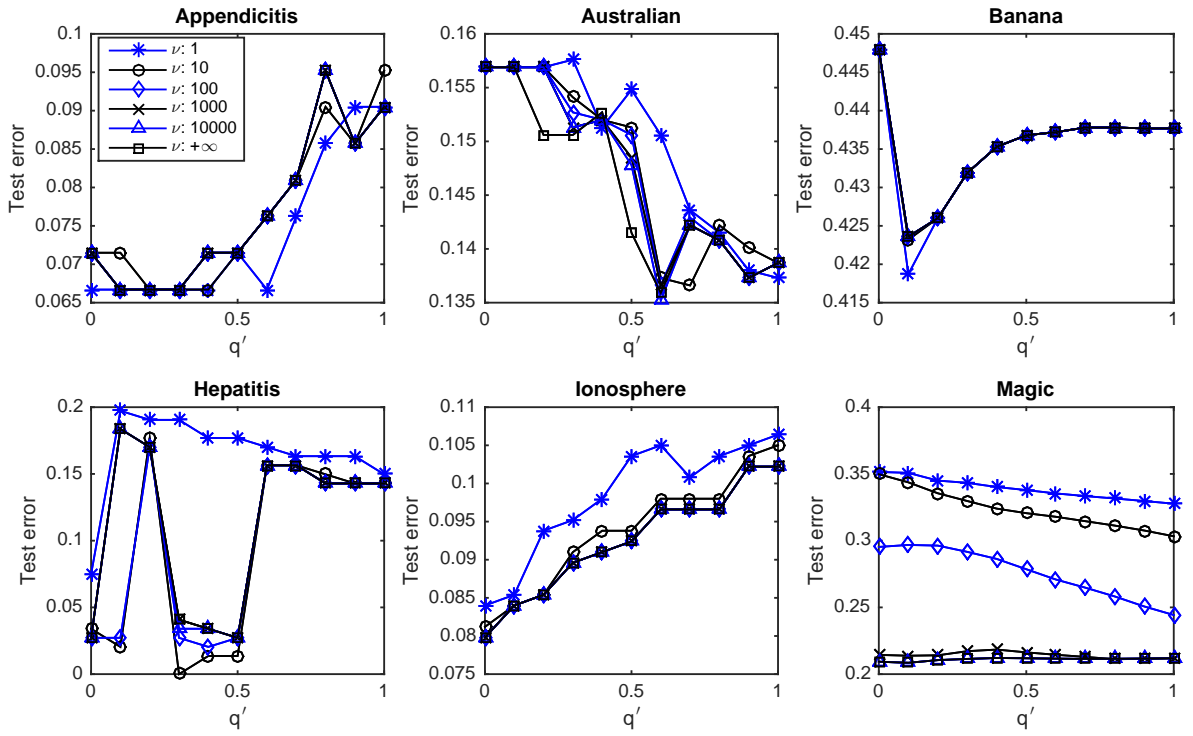


Fig. 1. Test error vs. q' for binary classification task